



Response from the AI Risk and Vulnerability Alliance to the NTIA AI Accountability Policy Request for Comment

*Carol Anderson (carol@avidml.org), Machine Learning Lead, AI Vulnerability Database
Borhane Blili-Hamelin (borhane@avidml.org), Secretary, AI Risk and Vulnerability Alliance
Subhabrata Majumdar (subho@avidml.org), President, AI Risk and Vulnerability Alliance
Nathan Butters (nathan@avidml.org), Founding Director, AI Risk and Vulnerability Alliance*

We are ARVA, the AI Risk and Vulnerability Alliance, a 501(c)(3) nonprofit. Our mission is to empower communities to recognize, diagnose, and manage vulnerabilities in AI systems. Our flagship project is [AVID](#), the AI Vulnerability Database. Inspired by best practices from cybersecurity, AVID combines an authoritative taxonomy of AI risks with an open-source, actionable knowledge base of flaws in AI systems and known mitigation techniques.

In our response to NTIA's request for comments, we address two questions closely related to our mission. In our answer to question 11, we discuss lessons from cybersecurity that can be applied in the AI accountability ecosystem. In response to question 23, we discuss the need for centralized and standardized reporting of AI accountability "products", focusing in particular on the reporting of vulnerabilities.

11. What lessons can be learned from accountability processes and policies in cybersecurity, privacy, finance, or other areas?

Cybersecurity researchers and practitioners have long recognized the need to document and share information about known security flaws and weaknesses — vulnerabilities. To facilitate communication and collaboration, standardized documentation and centralized databases of threats and failures have been developed. These include:

- The Common Vulnerabilities and Exposures (CVE) system, which provides a standardized naming scheme for vulnerabilities

- The Common Vulnerability Scoring System (CVSS), which is used to provide qualitative assessments of severity
- Centralized, public databases of vulnerabilities and exploits, including NIST’s National Vulnerability Database (NVD), the Open Source Vulnerability Database (OSVDB), MITRE ATT&CK, & MITRE ATLAS, which contain technical details and remediation techniques

AI systems are prone to failures that go far beyond security (vulnerability to intentional exploits) to include silent, unintentional failures pertaining to human rights, discrimination, reliability, measurability, transparency, and misuse. The breadth, explosive growth, and socio-technical complexity of new AI tools all too often prevents teams from learning from the mistakes that have already been made by other AI builders. Very few centralized resources currently exist to allow practitioners to efficiently and unambiguously exchange information about weaknesses and failures in AI systems.

We believe that the resources listed above, developed in the cybersecurity field to enable effective information sharing, should serve as models for developing similar resources in the field of AI. To that end, we’re developing the AI Vulnerability Database (AVID) as an open-sourced, community driven repository for the disclosure of social and technical vulnerabilities. Here we use “vulnerability” to refer to any weakness in an AI system that has the potential to result in an incident; this can include bias, poor accuracy, and other issues noted above.

We see the following features as important components of an open-source database for AI vulnerabilities:

- **A comprehensive taxonomy of AI risks** to enable classification of vulnerabilities along both social and technical dimensions.
- **Incident and vulnerability reports** that enable follow-up fixes and risk assessment of future similar workflows.
- **Severity scoring** to enable risk assessment and prioritization of mitigation efforts.
- **Remediation techniques**, when available.

A flexible, standard technical infrastructure operationalizes all of the above, through a common classification of risk and mitigation categories and schematization of incident and vulnerability reports. This enables any organization developing AI to interface with and build upon our database, empowering their engineers, data scientists, and risk management professionals to proactively look out for and fix the failure modes of the AI they develop.

Another lesson from NIST NVD, CVE, and CVSS is the importance of a robust adjudication process for AI vulnerabilities. At AVID, we see this problem as, on the one hand, a standard

setting effort. There are currently no broadly accepted and transparent criteria for AI vulnerabilities and severity scoring, and for many areas of mitigation best practices. Establishing a trusted adjudication process requires multi-stakeholder standard setting with buy-in from cross-sector stakeholders such as partner organizations, AI risk management professionals, AI builders, and impacted communities. On the other hand, trusted adjudication requires a robust editorial process. Every entry needs to be adequately vetted. The editorial process needs to be capable of vetting large numbers of quality entries. Finally, the editorial process needs to be accountable and transparent to outside parties, in a way that enables meaningful contestation.

23. How should AI accountability “products” (e.g., audit results) be communicated to different stakeholders? Should there be standardized reporting within a sector and/or across sectors? How should the translational work of communicating AI accountability results to affected people and communities be done and supported?

Here, we focus our answer on the reporting of vulnerabilities (known flaws or weaknesses) in AI systems. AI audits, of course, may also cover less-technical aspects of a system such as accountability structures, documentation, and compliance with relevant legal frameworks, and these are just as critical to consider.

We strongly support the development of standardized reporting formats and centralized repositories for AI vulnerabilities. In cybersecurity, vulnerability reports adhering to the standardized CVE format have proven highly effective at facilitating communication among practitioners. Drawing inspiration from CVE reports, at AVID we have developed a standard format for AI vulnerability reports and a standardized, public database to collect such reports. The availability of this information in a central repository is crucial to allow researchers and practitioners to benefit from each other's work. Standardized formats allow interoperability between federated databases collecting these reports. Standardization is also crucial for enabling researchers to aggregate and analyze reports from multiple sources.

For such reports to support accountability, disclosure of results (e.g. vulnerabilities, audit results) to the public must navigate the competing needs of the public and organizations. Whether, when, and how public disclosures are made should be standardized across sectors to the extent possible. While the risks of disclosures may vary in specific use cases, broad consistency in the disclosure process is vital. The concept of ethical disclosure—reporting the vulnerability or results to the company before the public—should be used where there is evidence that the reports will be taken seriously and the problem mitigated. Full disclosure—reporting the vulnerability or results to the public without warning the

company—should go through an adjudicating body informed by the relevant ethical considerations. Disclosures are needed to incentivize fixing the flaws and to empower other organizations to avoid making the same mistakes. Yet as we know from cybersecurity disclosures, informing the public immediately carries the danger of further harm: from enabling adversarial actions to perversely de-incentivizing knowledge sharing and transparency.

The adjudicating body should be responsible for, and therefore requires material support to do, the translational work of taking technical reports or socio-technical audits and bringing them to affected communities in a meaningful way. We believe that there are three specific outcomes that this translational work should achieve:

- Informing communities - make the people affected by the AI system in question aware that they can be targeted or impacted by the vulnerability.
- Preventing future harm - provide access to developers of systems in a way that allows them to mitigate the vulnerability or choose safer options for their systems.
- Enabling contestation and dialogue - bringing people together to discuss how the AI is being used and who it benefits to offer recourse to those negatively affected and address inequalities in the way benefits are distributed by the use of AI.

Conclusion

At ARVA, we want to build a world in which the resources that inform the daily work of practitioners are co-created by the many communities working to make AI less harmful. We are thankful for this opportunity to provide input on the NTIA's request for comments. Please do not hesitate to reach out if we may be of further help.